PR #21534 完整报告

sgl-project/sglang

[AMD] Add GLM-4.7-FP8 accuracy CI test for MI35x

合并时间：2026-03-28 15:28

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21534

执行摘要

该 PR 为 AMD MI35x 硬件添加了 GLM-4.7-FP8 模型的 GSM8K 精度 CI 测试，通过新增测试文件和 CI 任务扩展了测试覆盖，确保模型在特定配置下的准确性。

功能与动机

动机是添加 GLM-4.7-FP8 模型在 AMD MI35x 上的夜间 GSM8K 精度测试，以验证模型在该硬件上的性能。PR body 明确指出目标是“Add nightly GSM8K accuracy test for GLM-4.7-FP8 on AMD MI35x (8-GPU TP8).”

实现拆解

- 测试文件：新增 test/registered/amd/accuracy/mi35x/test_glm47_fp8_eval_mi35x.py，定义了一个使用 GSM8K 数据集、TP8 配置和特定解析器的精度测试类 TestGLM47FP8EvalMI35x。
- CI 配置：修改 .github/workflows/nightly-test-amd-rocm720.yml，添加新的 CI 任务 nightly-8-gpu-mi35x-glm47-fp8-rocm720，配置在 MI35x 硬件上运行上述测试。

评论区精华

- 重复参数设置：reviewer gemini-code-assist[bot] 指出测试中重复设置 --tp=8 参数，可能导致启动错误。作者 Jacob0226 已修复。
- 硬编码路径：reviewer 批评硬编码 Hugging Face 缓存路径，但作者辩称是 CI 环境特定需求。
- 代码重复：reviewer 建议抽象共享逻辑，作者认为当前代码足够简洁，暂不实施。

风险与影响

风险：重复参数可能引起模型启动异常；硬编码路径降低测试可移植性；代码重复增加维护负担。

影响：对系统增加 CI 负载，但对用户提升了模型可靠性和对团队增强了测试能力。

关联脉络

与此 PR 相关的历史 PR 包括其他 CI 和测试改进，如 PR 21585 和 21579，但本 PR 专注于 AMD 硬件特定测试，是测试套件扩展的一部分。